Scientific and Technical Report

Sponsored by
Advanced Research Projects Agency/ITO
and United States Patent and Trademark Office

Browsing, Discovery and Search in Large Distributed Databases
of Complex and Scanned Documents

ARPA Order No. D570

Issued by EXC/AXS under Contract #F19628-95-C-0235

Date Submitted: July 9, 1998

Period of Report: April 1, 1998 to June 30, 1998

Submitted by: Professor W. Bruce Croft, Principal Investigator
Computer Science Department
University of Massachusetts, Amherst

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Advanced Research Projects Agency or the U.S. Government.

Distribution Statement A: Approved for public release; distribution is unlimited.

DTIC QUALITY INSPECTED 1

19980717 016

UNIVERSITY OF MASSACHUSETTS
AMHERST

Computer Science

Lederle Graduate Research Center
Box 34610
Amherst, MA 01003-4610
(413) 545-2744

DATE: July 9, 1998

TO: Defense Technical Information Center (DTIC)

FROM: W. Bruce Croft, Principal Investigator

SUBJECT: Quarterly Scientific and Technical Report for F19628-95-C-0235

Enclosed is your required number of copies of the quarterly Scientific and Technical Report for ARPA order number D570 (note: changed from old AO #D468) issued by ESC/AXS under contract number F19628-95-C-0235. The R&D Status Report will follow by the due date. The title of the project is "Browsing, Discovery and Search in Large Distributed Databases of Complex and Scanned Documents." These reports are being distributed in the appropriate amounts to ESC/AXS, ESC/ENK, ARPA/ITO, DTIC, and ARPA/Technical Library.

If you have any questions, I can be reached by email at croft@cs.umass.edu.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

1\. AGENCY USE ONLY *(Leave blank)*

2\. REPORT DATE
07/09/98

3\. REPORT TYPE AND DATES COVERED
Scientific/Tech 04/01/98 - 06/30/98

4\. TITLE AND SUBTITLE
Browsing, Discovery, and Search in Large Distributed
Databases of Complex and Scanned Documents

5\. FUNDING NUMBERS
F19628-95-C-0235
ARPA Order No. D570

6\. AUTHOR(S)
W. Bruce Croft

7\. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Massachusetts, Amherst
Box 36010, OGCA, Munson Hall
Amherst, MA 01003-6010

8\. PERFORMING ORGANIZATION REPORT NUMBER
TR5281810798

9\. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Mr. Harry Koch
ESC/AXS
Bldg 1704. Room 114
5 Eglin St.
Hanscom AFB, MA 01731-2116

Ms. Monique Dillon
Office of Naval Research
Boston Regional Office
495 Summer St., Room 103
Boston, MA 02210-2109

10\. SPONSORING/MONITORING AGENCY REPORT NUMBER

11\. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution Statement A: Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13\. ABSTRACT *(Maximum 200 words)*

This project aims to integrate powerful, new techniques for interactive browsing, discovery, and retrieval in very large, distributed databases of complex and scanned documents. Emphasis is placed on going beyond full-text retrieval techniques developed in the DARPA TIPSTER program to support different types of access and non-textual content. These techniques should be particularly relevant to the patent domain where it is important to find relationships between documents and where the patent or trademark may be based on a visual design. The specific tasks identified involve studying representation techniques for long documents with complex structure, browsing and discovery techniques for large text databases, image retrieval and scanned document retrieval techniques, and architectures for large, distributed databases.

14\. SUBJECT TERMS

Browsing	Query Processing	Indexing
Image Retrieval	Scanned Document Retrieval	Bayesian Network
Text Retrieval	Probabilistic Retrieval Model	Large Distributed Databases

15\. NUMBER OF PAGES
10

16\. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

Browsing, Discovery and Search in Large Distributed Databases of Complex and Scanned Documents

Technical and Scientific Report

Task 1: Representation Techniques for Complex Documents

Task Objectives

In this task, the goal is to extend the word-based representations that are common in retrieval systems in order to support summarization, browsing, and more effective retrieval. Specifically, we will be studying phrase-based representations and relationships between phrases in individual and groups of documents as the basis for our approach. Document structure will be used as part of the information that is used to "tag" the phrasal representation.

Technical Problems

The technical problems have to do with defining a "phrase", developing techniques for rapidly extracting them from text, comparing phrase contexts to identify significant relationships, producing summaries from these representations, extending the underlying retrieval model to be able to make effective use of phrasal representations, and using complex document structure in indexing and retrieval.

General Methodology

The general methodology for this task is to demonstrate effectiveness through user-based and collection-based experiments. As well as the PTO text databases, we will make extensive use of the TIPSTER document collection, which consists of a large number of text documents from a variety of sources, queries, and user relevance judgments for each query.

Technical Results

We continue to develop phrase-based techniques for the patent retrieval system. Work on the phrase recognizer is progressing, although there have been problem with the recall levels of this approach. We are improving the training data to address this issue.

As a result of a visit to the PTO in May, we have begun work on a tool for phrase displays to incorporate into the patent searching process. The idea of this tool is to allow examiners to look at phrases from specific classes that are related to query words and phrases. The first version of this tool shows phrases that are modifications of query words and phrases that are associated with query words by significant co-occurrence. For example, in the speech recognition class, the tools find the following list of modifications for "phone" (this is a subset of the total list):

```
cellular phone
composite phone
corresponding phone
fenemic phone
phone baseform
phone call
phone change
phone choice
phone classes
phone context
phone directory
phone label
phone levels
phone line
phone lines
phone machine
phone machines
phone models
phone number
phone numbers
phone operation
phone sequence
phone weights
```

The same tools find the following associations for "phone":

```
phone
machines
phones
phone machine
cellular
manufacturing
markov
networks
markov models
touch
machine
phone number
touch screen
speaking modes
trivial task
enormous problems
trivial
voice-dial cellular phone
speech-recognition devices
manufacturing process
automated speech-recognition device
natural interface
background noise present
prosthetic
speaker variability
machine interface
label
query
automated speech recognition
speaking speed
models of words
```

It can be seen from this list that, although the words and phrases found are generally useful, more work needs to be done to filter out general words and phrases such as "trivial", and "enormous problems".

Important Findings and Conclusions

Initial experiments with the new phrase recognition approach showed high precision but relatively low recall. To be useful in practice, the recall level must be improved and we are currently working on this.

Significant Hardware Development

None

Special Comments

None.

Implication for Further Research

We plan further development and testing of the phrase recognition approach. We also will continue to incorporate the new phrase tools into a patent search demonstration.

Task 2: Browsing and Classification Techniques for Document Collections

Task Objectives

The goals of this task are to develop techniques for summarizing and classifying collections of documents. These techniques will be designed to support interactive browsing and text classification in environments like the PTO.

Technical Problems

The technical problems involve producing an effective summary of a group of documents, such as a retrieved set or an entire database. Both document and phrase clusters could be used as part of this process. The classification task emphasizes the ability to accurately assign predefined categories (as in the PTO classification) to new documents (patents). An additional problem is to determine when existing classifications do not match well to new documents, such as when a PTO category covers too many patents and needs to be refined.

General Methodology

Evaluation of these techniques will be done using both the TREC corpus and PTO data. For the classification task in particular, we are designing evaluation criteria with substantial input from PTO staff.

Technical Results

A number of classification experiments were performed on the new data from the speech patent classes. This data contains all the documents from 1985-1997 in the roughly 100 subclasses under the "speech signal processing" node in the PTO hierarchy. We have been training classifiers to place documents in the 30 or so subclasses fall under the "speech recognition" node, using the others as closely related negative examples. The years 1985-1995 has been used as training data and 1996-1997 as test data. The focus of this work has been to study if the hierarchical structure of the patent classes could be used to improve the document placement accuracy. This accuracy was very high for the initial task specified by the PTO (94% correct), but the speech task is much more difficult in that the subclasses are much more closely related and overlapping.

A revised version of the paper was prepared for the AAAI Workshop on Learning for Text Categorization, entitled “Some issues in the automatic classification of US Patents.”

Experiments on the visualization of retrieval results continue. One new paper was prepared.

Leouski, A. and Allan, J., "Evaluating a Visual Navigation System for a Digital Library," submitted to the Second European Conference on Research and Technology for Digital Libraries, 21-23 September, Heraklion, Crete, Greece.

Important Findings and Conclusions

Our experiments so far have not been able to improve accuracy using the hierarchical patent structure. The document placement accuracy in the speech area is significantly lower than in the much simpler early test, but this is still being worked on. During this work, a problem was identified that has subsequently been confirmed with the PTO. The problem is that the PTO has misclassified a number of the patents in the speech area. This creates significant difficulties for evaluating our automatic classification algorithms.

Significant Hardware Development

None

Special Comments

None.

Implication for Further Research

We continue to focus on improving the classification accuracy, incorporating additional classification techniques into the classification system, and evaluating visualization techniques.

Task 3: Image Indexing and Retrieval

Task Objectives

The goal of this task is to develop similarity-based techniques for retrieving images such as trademarks, logos, and designs.

Technical Problems

The central issue is how images can be indexed to support efficient, content-based retrieval. The primary type of query in these environments is “find me things that look like this”. We are developing "appearance-based" retrieval of images as well as more straightforward features such as color and texture. Filter based and frequency domain based techniques offer some potential in this area, but significant work needs to be done on making this approach efficient enough to deal with hundreds of thousands of images.

General Methodology

The evaluation of these techniques will be done in a similar way to text by developing test collections of images. Specifically, we are working to obtain large collections of trademark and design images, both from the PTO and from general sources such as the web.

Technical Results

We have continued to improve our trademark appearance-based retrieval techniques. Currently, our work is focused on improving the precision for the 60,000 trademark database. We are experimenting with increasing the amount of spatial information used for indexing and retrieval by partitioning the histograms used. Each change involves recomputing some of the basic features and indices. Due to the size of the database involved, each change takes a considerable amount of time to process. Preliminary indications are that the precision is increased significantly by including the extra spatial information.

We have also increased the efficiency of our code so that the server takes up less memory. Other work has involved removing some of the bugs in the interface. Many of these bugs arise because even the same browser (eg Netscape) seems to behave differently when the operating system is different.

We have also completed the first revision of the 650,000 trademark database, which involved resampling all images to full and thumb images, recomputing JPEG features, and modifying the interface to support multi-resolution searches.

We have developed a demonstration system based on detecting flowers in plant patents. The demonstration consists of an interface that can retrieve flowers of a given color. The query can be presented in the form of a color name or, alternatively, an example picture of a flower of a given color may be used. During the indexing process, the picture is segmented to find the flowers separately from the background using domain knowledge. The colors of the flower are then stored in the index. The database currently contains 70 plant patents and about 230 other images takes from a variety of sources. We have asked the PTO for more flower patent images.

Based on a discussion with the PTO, we have also begun to examine design patents for fonts as a potential application of image matching.

The following papers that were recently submitted have been accepted for presentation.

1) Ravela, S. and Manmatha, R., "On computing global similarity in images" accepted to the IEEE Workshop on Applications of Computer Vision (WACV"98), October 1998.

2) Das, M., Manmatha, R. and Riseman, E. M., "Indexing Flowers by Color Names using Domain Knowledge-driven Segmentation" accepted to the IEEE Workshop on Applications of Computer Vision (WACV"98), October 1998.

Important Findings and Conclusions

The inclusion of more spatial information appears to improve the accuracy of trademark matching. Flower retrieval by color name works well, but we have no feedback yet on whether this will be useful for patent examiners.

Significant Hardware Development

None

Special Comments

The progress of this part of the project depends on data from the PTO. Specifically, we still need design code data for the full trademark database and more flower patents.

Implication for Further Research

We will continue to evaluate the accuracy of trademark and flower retrieval, and develop new design patent applications.

Task 4: Distributed Retrieval Architecture

Task Objectives

The goals of this task are to scale up our current methods of automatically selecting collections and merging results, and to investigate architectures that can support efficient retrieval, browsing and relevance feedback in distributed environments with terabytes of information.

Technical Problems

The current INQUERY text retrieval system uses a client server architecture to support simultaneous retrieval from multiple collections distributed across one or more processors. A number of efficiency bottlenecks develop, however, when the size of the databases is very large. Deciding which subcollections to search can address part of the problem, but there are other problems associated with the fundamental efficiency of the processes involved and the use of distributed resources. Image indexing and retrieval tends to make all of these problems worse since the databases and indexes are considerably larger.

General Methodology

The architectures and algorithms produced in this task will be evaluated using a combination of standard performance (efficiency) measures and effectiveness measures. The efficiency tests will be done using TREC data and large PTO databases, including images, and the collection selection algorithms will be evaluated using the text subcollections of the patents.

Technical Results

We continued to develop the distributed testbed using the fast network until it was shut down. We are now supporting the demonstration using a regular T1 link, which is considerably slower.

We have begun to evaluate the use of replicas in a distributed environment as a means of improving the scalability of retrieval systems.

Papers related to this effort:

- (1998), Lu, Z. and McKinley, K., “Partial Replica Selection based on relevance for Information Retrieval”, CIIR technical report.
- Lu, Z., McKinley, K. and Cahoon, B., “The Hardware/Software Balancing Act for Information Retrieval on Symmetric Multiprocessors”, to appear in the Proceedings of EuroPar98, Southampton, U.K. on Sept.1-4 , 1998, UMass technical report TR98-25

Important Findings and Conclusions

The distributed architecture used in InQuery was successfully used over the fast network. This experiment highlighted some important performance issues that need to be addressed.

Significant Hardware Development

None.

Special Comments

None

Implications for Further Research

We will continue to evaluate performance of distributed architectures for scalable IR.